Intro to Humanities Data

Simple Visualizations for Complex Arguments

Stacey Giroux
Assistant Research Scientist, Center for Survey Research and Department of Anthropology
sagiroux@indiana.edu

Mia Partlow, @mia_partlow
Digital Humanities Fellow, Information & Library Science
Digital Methods Specialist, Institute for Digital Arts & Humanities
mapartlo@indiana.edu

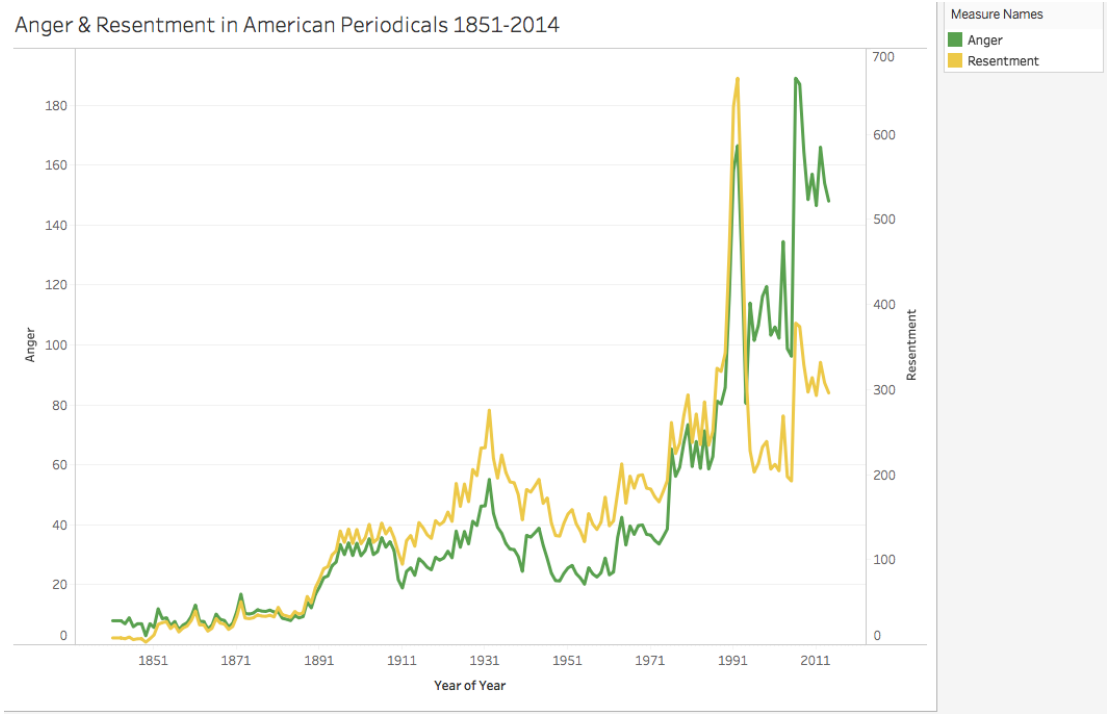
What We're Doing Today

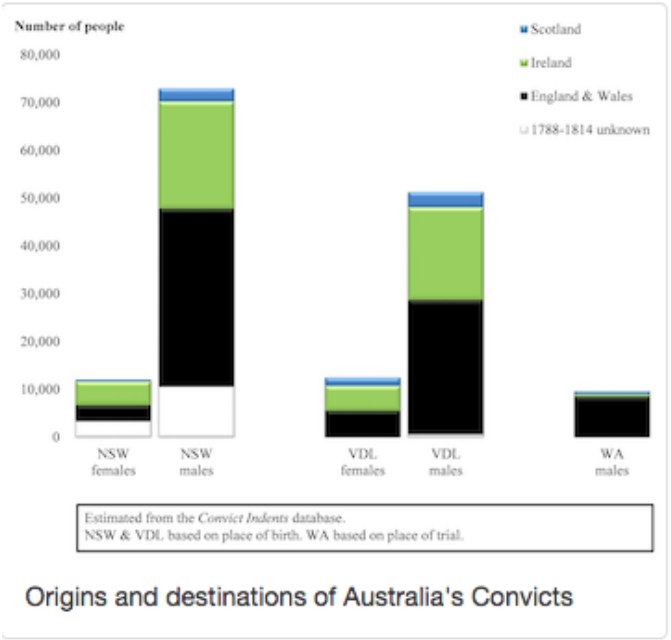
- A brief review of Humanities Data
- Why we need to learn statistical concepts
- Our two sample datasets
- Statistical concepts and useful vocabulary
- Hands-on visualization of our data
- Hands-on statistical test with our data!

Humanities Data

- Data are gathered and created as part of the research process and are often:
  - Interpretive
  - Ambiguous
  - Unstructured
- Data need to be structured to meet the needs of your visualization method or statistical test
- Your research question will ground the process of creating and structuring your data

Humanities Data Visualization





Source: https://www.digitalpanopticon.org

Sample Dataset: Tracking Word Usage

- Research Project: tracking use of the word “resentment” in American periodicals.
- Data source: ProQuest Historical Newspapers Database
- Issues:
  - Average per year not contextualized within number of words per issue

Year	Chicago Defender	Chicago Tribune	Los Angeles Sentinel	Los Angeles Times	New York Times	New York Tribune	Grand Total	Average Mentions
1841						8	8	8
1842						8	8	8
1843						8	8	8
1844						7	7	7
1845						9	9	9
1846						6	6	6
1847						7	7	7
1848						7	7	7
1849						3	3	3
1850						7	7	7
1851					7	16	23	11.5
1852					37	11	48	24
1853					43	9	52	26
1854					44	10	54	27
1855					30	8	38	19
1856		8			37	25	70	23.33333333
1857		3			30	12	45	15
1858		7			32	20	59	19.66666667
1859		9			42	15	66	22
1860		8			54	23	85	28.33333333

Sample Dataset: 1880s Children's Literature

- Research Question: Is there a significant difference in the number of words used by male & female authors?
  - Hypothesis: e.g., women, on average, will have written shorter books published during this time period than men.
- Data: a corpus of children's literature and the associated metadata
- Metadata gives us the author's gender and the number of words per publication

	B	G	H	I	J	K	L	M
1	title	author last name	author gender	author nation	genre	source	year	words
2	A Dog with a Bad Name	Reed	Male	England	novel	Project Gutenberg	1886	96946
3	A Final Reckoning	Henty	Male	England	novel	Project Gutenberg	1887	100724
4	A House Party, Don Gesualdo, and A Rainy June		Female	England	short stories	Project Gutenberg	1887	85407
5	A Houseful of Girls	Tytler	Female	Scotland	novel	Project Gutenberg	1889	93410
6	A Little Country Girl	Coolidge	Female	America	novel	Project Gutenberg	1885	48458
7	A Round Dozen	Coolidge	Female	America	novel	Project Gutenberg	1883	47721
8	A Sailor's Lass	Leslie	Female	England	short story	Project Gutenberg	1886	22292
9	A World of Girls	Meade	Female	Ireland/England	novel	Project Gutenberg	1886	82057
10	Adrift in the Wild	Ellis	Male	America	novel	Project Gutenberg	1887	62594
11	Adventures in Africa	Kingston	Male	England	novel	Project Gutenberg	1883	36295
12	Adventures in Australia	Kingston	Male	England	novel	Project Gutenberg	1885	36160
13	All Adrift	Optic	Male	America	novel	Project Gutenberg	1882	65662
14	Battles With the Sea	Ballantyne	Male	Scotland	novel	Project Gutenberg	1883	27042
15	Bimbi: Stories for Children		Female	England	short stories	Project Gutenberg	1882	44327

(Source: Alan Liu, ed. Digital Humanities Resources for Project Building. http://dhresourcesforprojectbuilding.pbworks.com/)

Samples and Related Considerations

Target population and generalization

What's the population you want to be able to say something about, to make inference about?

Sampling frame

Materials, procedures, and devices (lists, maps) that identify, distinguish, and allow access to the elements of the target population. (Lessler and Kalsbeek 1992:44)

Undercoverage

Ineligibles

Sample Types

Probability sample

Census

Random sample, stratified random sample, etc.

Non-probability sample

Convenience sample, snowball sample, etc.

Example: Resentment Dataset

Sampling frame

ProQuest Historical Newspapers database

Sample type

Non-probability (purposive) sample of newspapers from that database

Generalizability of sample

These three papers, this time period

Titles Included

ProQuest Historical Newspapers™ is the definitive news
Century.

Featured Titles

- *The Arizona Republican*—1890-2007
- *Atlanta Constitution*—1868-1984
- *Austin American Statesman*—1871-1978*
- *The Baltimore Sun*—1837-1991*
- *The Boston Globe*—1872-1985*
- *The Chicago Tribune*—1849-1993*
- *The Christian Science Monitor*—1908-2004*
- *Cincinnati Enquirer*—1841-2009
- *Dayton Daily News*—1898-1922
- *Detroit Free Press*—1831-1999
- *Hartford Courant*—1764-1991*
- *Indianapolis Star*—1903-2004
- *Los Angeles Times*—1881-1993*
- *The Louisville Courier-Journal*—1830-2000
- *Minneapolis Star Tribune*—1867-2001
- *Nashville Tennessean*—1812-2002
- *The New York Times* (1851-2014*) *with Index* —(1851-1
- *New York Tribune / Herald Tribune*—1841-1962
- *Newsday*—1940-1989*
- *The Philadelphia Inquirer*—1860-2001
- *Pittsburgh Post-Gazette*—1786-2003
- *San Francisco Chronicle*—1865-1922
- *St. Louis Post-Dispatch*—1874-2003
- *Wall Street Journal*—1889-2000*
- *Washington Post*—1877-2000*

International Newspapers

- *Chinese Newspapers Collection*—1832-1953

Example: Children's Literature Dataset

Sampling frame

Project Gutenberg

Sample type

Census? Random?

Generalizability

Only to those texts

Data Types

Cross-sectional

	A	B	H	I	J	K
1	ID	title	author ger	author natio	genre	source
2	abirdschristmascarol	The Bird's Christmas Carol	Female	America	short story	Project Gutenberg
3	thestoryofpatsy	The Story of Patsy	Female	America	short story	Project Gutenberg
4	thebeemanoforn	The Bee-Man of Orn	Male	America	short storie	Project Gutenberg
5	alittlecountrygirl	A Little Country Girl	Female	America	novel	Project Gutenberg
6	arounddozen	A Round Dozen	Female	America	novel	Project Gutenberg
7	adriftinthewilds	Adrift in the Wild	Male	America	novel	Project Gutenberg

Panel

Time series

Publication	Year	Word	Resent_Mentions
Chicago Tribune	1856	Resentment	8
Chicago Tribune	1857	Resentment	3
Chicago Tribune	1858	Resentment	7
Chicago Tribune	1859	Resentment	9
Chicago Tribune	1860	Resentment	8
Chicago Tribune	1861	Resentment	16

Variable Types

Categorical

title	author gender	author nation	genre
Heidi	Female	Switzerland	novel
Johnny Nut and the Golden Goose	Male	Scotland	short story
Prince Prigio	Male	Scotland	short story

Scale

genre	source	words
novel	Project Gutenberg	63660
novel	Project Gutenberg	62594
short story	Project Gutenberg	55822
novel	Project Gutenberg	100909

Publication	Year	Word	Resent_Mentions
Chicago Tribune	1856	Resentment	8
Chicago Tribune	1857	Resentment	3
Chicago Tribune	1858	Resentment	7
Chicago Tribune	1859	Resentment	9

Data Types, Variable Types, & Statistical Testing

Cross-sectional data (children's literature)

- Categorical variables: Is there an association between author's sex and genre of the work composed?

  Chi-square test: test association between two categorical variables

- Scale: Is there a significant (i.e., statistically significant) difference in the word length of books between men and women?

  Independent samples t-test: can test difference between means

  Null hypothesis: There is no significant difference in book length

Time-series data (anger/resentment data)

Special methods

Assumptions of these tests: *chi-square*

- Two categorical variables
- Observed frequencies, not rates or percentages
- Independent observations
- Random or stratified random sample
- Large enough sample size
- Normally distributed data

There would be other details to attend to with this dataset

Assumptions of these tests: *t-test*

For testing whether the difference between two values is significant, that is, greater than we would expect by chance

- Dependent variable is continuous
- Random or stratified random sample
- Large sample
- Normally distributed data
- No significant outliers

Hands-on Data Visualization

Download the spreadsheet and open it in Excel:

http://tiny.cc/resentment-workshop

-and-

Go to:

http://tiny.cc/color-simulator

T-Test

To follow along, download the data and instructions here:

http://tiny.cc/t-test

Resources

Campus Resources

- Center for Survey Research
  - Email Stacey: sagiroux@indiana.edu
- Social Science Research Commons
  - http:ssrc.indiana.edu
- IDAH Consultation Hours
  - http://idah.indiana.edu/project-support for hours
  - Or email idah@indiana.edu
- Online Tutorials
  - https://statistics.laerd.com

Upcoming Events:

Intro to Humanities Data: The Path to Complex Visualizations and Statistics	12pm, January 30, Hazelbaker Hall
Virtual Realities: Making the Humanities in a Digital World, a talk by Dr. William "Bro" Adams	5:30p, February 20, Woodburn Hall 100